Meeting Summary
Final Version
Page One Editorial Services
Last Updated: April 25, 2001

Summary of Workshop on
Bioinformatics Strategies for Application of Genomic Tools to Environmental Health Research

March 5, 2001
The McKimmon Center
North Carolina State University, Raleigh, North Carolina

Introduction

The National Institute of Environmental Health Sciences (NIEHS) established the National Center for Toxicogenomics (NCT) in June 2000. Toxicogenomics is an emerging scientific field that combines studies of genetics, genome-wide mRNA expression, cell and tissue-wide protein expression, and bioinformatics to understand the roles of gene-environment interactions in disease. The NCT was created to facilitate application of toxicogenomics to improve human health.

The goals of the NCT are: 1) to develop and apply gene expression and proteomics technology to study the biological effects of chemicals and drugs; 2) to support intramural and extramural research to define the effects of environmental agents on gene expression; and 3) to develop a national reference and relational database on "Chemical Exposure in Biological Systems" (CEBS) that will serve as a resource in the fields of toxicology and environmental health.

To help define the path towards fulfilling its goals, a series of three workshops were held through which the research community could learn about the NCT and provide input to the NCT during its early stages. The first workshop in this series entitled "Functional Genomics and Environmental Health" was held at Massachusetts Institute of Technology in Cambridge, Massachusetts on December 11, 2000, and the second workshop entitled "Functional Proteomics in Environmental Health Science" was held on January 24, 2001 at the University of Arizona in Tucson, Arizona (reports of these meetings are available at http://www.niehs.nih.gov/nct/workshop.htm). This meeting summary reports on the third workshop in the series entitled "Bioinformatics Strategies for Application of Genomic Tools to Environmental Health Research" which was held at North Carolina State University (NCSU) on March 5, 2001. The meeting was co-organized by Bruce Weir (NCSU), Cynthia Afshari (NIEHS) and Pierre Bushel (NIEHS) and co-sponsored by NIEHS and NCSU.

In his introductory comments, Bruce Weir expressed enthusiasm for the potential of emerging genomics-based technologies and for programs such as the NCT, which hope to develop these technologies to benefit human health. Bioinformatics is essential to this effort and NCSU is

establishing a new Bioinformatics Research Center (officially opening on March 6, 2001), which will be one of the first degree granting programs in Bioinformatics in the Nation. NCSU has an ongoing close partnership with NIEHS and hopes to continue to build on this interaction as part of the NCT.

Analysis of Global Patterns of Gene Expression

John Quackenbush (The Institute for Genomic Research) discussed methods for analysis of large amounts of DNA sequence and gene expression data. He emphasized that this effort should be seen in a historical context. With the advent of high throughput DNA sequencing technology, the twentieth century brought us into the " golden age of genetics." At least 30 microbial genomes have been sequenced and approximately 50 more are in the process of being sequenced. Sequences of eukaryotic genomes are also being completed rapidly, including genomes of yeast, *Caenorhabditis elegans*, *Drosophila* and *Arabidopsis*. A working draft of the human genome is complete and mouse and rat sequencing projects are nearing completion. The expressed sequence tag database (dbEST) has at least 6 million entries, and almost half of these are from the human genome. The rate of progress is increasing, the cost of sequencing is decreasing and new technologies are rapidly moving from the lab to the field. Because of these trends, biology is rapidly becoming an information science. Nevertheless, the biological framework remains critical to understanding and reaping benefit from the large amount of raw data and information being produced using genomics technology.

The challenge now facing biologists and biomedical scientists is to identify the genes, understand their function, discern biological roles of gene products and evaluate genes for being diagnostic or prognostic for disease. These goals can begin to be realized through comparative genomics. It is necessary to compile information and create database tools that cross index and compare large amounts of genomics information. These database tools will be most useful if they integrate different types of data including genomic sequences, transcript maps, expression data and gene function data.

The Institute for Genomics Research (TIGR) is taking on this challenge with the TIGR Gene Index Project. This effort began as a data mining problem to identify expressed DNA sequences, and grew into an effort to use sequence homology to identify orthologs and paralogs in the sequence database. The TIGR Gene Indices integrate data from international EST sequencing efforts and gene research projects creating a resource that collects and analyzes public EST data in a comprehensive manner. The raw data is usually "cleaned up" so that contaminating, redundant or undesired sequences are removed or masked (*i.e.*, vector DNA sequences). Sequences are compared using various clustering algorithms (*i.e.*, CAP3), a virtual transcript is created, and tentative consensus sequences (TCs) are derived. The TCs are then annotated and released. The TIGR Gene Indices have a user-friendly interface that provides homology search information and links to related biological information sources. The URL for the project is http://www.tigr.org/tdb/tgi.shtml.

TIGR is using ESTs to identify tentative orthologs in different species. It has established a TIGR Orthologous Gene Alignment (TOGA) database based on TCs represented in the TIGR Gene

Indices. TOGA currently is divided into separate sections for mammals, plants and parasites, and it will incorporate additional species as data become available. Groups of orthologous genes (TOGs) are identified by a reflexive transitive closure process involving sequential pairwise comparisons that search for a "best match" sequence. Both orthologs and paralogs can be identified by this process. In many cases, the searches are carried out using conserved 3'-UTR sequences, because these sequences are overrepresented in EST datasets. A similar approach is being used to identify conserved regulatory sequences.

TIGR has also developed several tools for analysis of microarray expression data. For example, TIGR developed "TIGR spotfinder," an image processing tool that identifies differentially expressed genes (*i.e.*, spotfinder draws a grid, calculates background and exports data to a spreadsheet or database). Quackenbush mentioned that there are several methods for normalization of microarray expression data (*i.e.*, total intensity, linear regression, ratio statistics and iterative log mean centering), and he emphasized that all these methods work; however, it is important to apply a method to the whole data set and it is better to use a combination of several methods than to rely on a single method.

Analysis of multiple experiments requires data mining, pattern recognition and clustering algorithms. It can be useful to use a geometric representation of the data with vectors whose coordinates are the log ratio values for each experiment. The number of experiments determines the number of dimensions required to represent the dataset, and distance metrics measure the distance between vectors. Many tools are available for these types of analyses such as hierarchical clustering, self-organizing maps and principal component analysis.

TIGR is currently involved in several projects using microarray technology. One such study is underway to develop an expression-based approach to classify tumors. Traditionally, tumors are classified based on their morphology, but this approach fails to reliably predict clinical outcome. It may be possible to identify expression fingerprints that predict the metastatic potential of a tumor. TIGR has analyzed expression patterns in tumor cell lines using a human DNA array for approximately 30,000 genes. Several potentially diagnostic genes have been identified. Additional studies are required to characterize these genes, but it is hoped that they may provide useful markers of metastatic potential. TIGR is also characterizing several mouse models for disease (*e.g.*, heart, lung, blood and sleep disorders) to understand diagnostic expression patterns associated with disease susceptibility and to explore gene-environment interactions. Another project at TIGR is to analyze all potential gene coding sequences on the second chromosome of *Arabidopsis*. This project is a comprehensive analysis that includes hypothetical genes identified by sequence analysis. Some of the hypothetical ORFs are differentially expressed under certain growth conditions.

A Public Gene Expression Data Repository

Alex Lash from the National Center for Biotechnology Information (NCBI) gave an overview of the NCBI Gene Expression Omnibus (GEO), a public gene expression data repository. GEO has 3 main goals: 1) to develop and maintain a durable repository for gene expression data; 2) to balance requirements for standards and flexibility; and 3) to provide this service without stifling

technological innovation. GEO is being developed as an online resource for retrieval of gene expression data from any organism or source.

GEO has a simple architecture with three data entities called **platform**, **sample** and **series**. **Platform** types include spotted glass slide arrays, high-density oligonucleotide arrays, hybridization filter arrays and serial analysis of gene expression. Data from any of these platforms will be accessioned to GEO and archived for public use. **Sample** information includes sample description and spot measurement data. **Series** describes the relationship between groups of samples (*i.e.*, dose response, time course, cell lines, strains, etc.). GEO imposes minimal restrictions on data that it accepts, because it does not want to exclude groups from submitting data. The sole platform restriction is that each spot is required to have a unique identifier. Other experimental information, including quality metrics, are optional. In addition, raw image data is not accessioned by GEO.

Data submission is through a user-friendly web-based interface (URL is http://www.ncbi.nlm.nih.gov/geo) . Tab delimited text tables are the primary data format. Accessions can be updated when additional information relevant to the dataset becomes available. Submitters can place a 6-month hold on release of accessioned data to allow for publication of their work. Data is retrieved by accession number or submitter list and can be downloaded. Current GEO holding include 2 platforms, 48 samples and 1 series.

Lash indicated that he is working on increasing the rate of submission to GEO. An Entrez-based interface is being developed to enhance user convenience, and work is underway to integrate GEO with other NCBI resources such as UniGene and LocusLink. In addition, Lash is collaborating with the Microarray Gene Expression Database (MGED) group on adopting the Microarray Markup Language (MAML) data exchange standard. Lash is also interested in developing a spot measurement database, which will allow data to be retrieved for a specific gene (clone/spot) of interest.

Statistical Design Issues in Microarray Experiments

Gary Churchill (The Jackson Laboratory) gave an overview of statistical design for microarray experiments. He opened his talk by quoting the statistician R. A. Fisher, who observed that "it is all too frequent for an experiment to be conducted so that no valid estimate of error is available." Churchill emphasized that this situation is unfortunate, because no conclusion can be drawn from such an experiment. In contrast, one can typically draw reasonable and valid conclusions from a well designed experiment that provides for valid estimation of error.

Good experimental design begins with defining several components and parameters including the following: 1) treatments to be applied and compared; 2) units that will be treated; 3) rules by which the treatments are allocated; and 4) specification of measurements. This experimental structure applies to microarray experimental design as well as other types of experiments. Data is collected from the array using a scanner and image analysis is carried out; image analysis is an important component of a microarray experiment, but it is not treated in the current discussion.

Churchill drew an analogy between statistical analysis of microarray data and analysis of data from agricultural experiments. In this analogy, a mRNA sample is equivalent to a crop variety and a spot on the array is analogous to a block of land. If 2 varieties are grown on 2 different blocks of land, variation may reflect characteristics of the variety and/or the block. To identify confounding effects, it is important to measure the yield of treated and untreated varieties on a single block and to measure the yield of some varieties on more than one block. If a higher number of blocks are tested, a higher number of replicates are made and a balanced experimental design is used, the statistical power of the experiment improves.

One of the simplest microarray experiments involves 2 treatments and 2 replicates; for example, mice are treated with and without drug, there are 2 mice per treatment group, and samples are labeled with 2 dyes and hybridized to 4 arrays. One can analyze this set of data in a daisy chain fashion using a closed loop of pairwise comparisons. Statistical analysis of the data can separate significant variation in the data from variation that is insignificant, uninformative or uninteresting.

Microarray slides have structural characteristics that depend on the characteristics of spots, genes, pingroups and subarrays. The spot is the smallest unit of the array. Measurement at an individual spot is a function of the characteristics of the array (A), dye (D), gene (G), variety (V) and spot (S). The investigator is usually least interested in variation due to A, D and S and most interested in the relationship between V and G. Thus, statistical approaches analyze variance caused by the parameters that are not of interest, and develop normalization methods that remove this variation.

Data is analyzed by applying techniques for analysis of variance (ANOVA). The main effects are due to A, D, or interactions between A and D or between A, D and V. However, there are also array by gene effects and dye by gene effects. Array by gene effects are variations due to spot size or spot quality which reflect the level of heterogeneity in the construction of the array. Dye by gene effects were unanticipated, but they are reproducible and a potential source of bias. Dye effects are best accounted for by a "flip-dye" experiment, in which differential dye-labeling is reversed with respect to sample; for example, experiment 1 is carried out by labeling sample A with dye1 and sample B with dye 2, and experiment 2 is carried out by labeling sample A with dye 2 and sample B with dye 1.

Churchill advocates a visual inspection of the data routinely used by Terrence Speed (see below) involving a plot of M vs. A, where $M = \log_2 [R/G]$, $A = \log_2 [(R*G)/2]$, with R = red fluorescence intensity, and G = green fluorescence intensity. There appears to be strong support for use of this approach to visualize statistical analysis of microarray data. The M vs. A plot evaluates the correlation between dye ratio and dye intensity across the array. Importantly, deviation from normality in this plot appears to be highest in the low intensity range.

Churchill described two data sets and demonstrated the usefulness of data normalization procedures. One data set, provided by the NIEHS Microarray Center, involved treated and control samples, triplicate replication, and fluor-flip duplication. Samples were hybridized to a total of 6 arrays. Churchill found that the dye by gene and gene by array effects were quite large prior to normalization, so that residual error was approximately 0.018. After use of a Lowess

curve normalization, residual error was reduced by an order of magnitude resulting in minimal dye by gene effects. Another approach that is possible in a highly replicated data set calculates a gene-specific error rate. He suggested that this can make it possible to detect a statistically significant change in gene expression as low as 1.2 fold. In reviewing the NIEHS data set and another large data set, Churchill emphasized that it is valuable and desirable to create a highly replicated set of data, because replication enhances the ability to make valuable statistical inference from the data.

The Second Simplest Microarray Data Analysis Problem

Terence Speed (University of California, Berkeley) indicated that the goal of statistical analysis is to separate systematic and random sources of variation in data. This is only possible if an experiment is properly controlled so that valid probability statements can be derived. Speed discussed "the simplest microarray data analysis problem, which he defined as follows: "is it possible to define a differentially expressed gene using one [microarray] slide?" According to Speed, this may be possible, however it is likely to require a process other than statistical inference (*i.e.*, visual examination of an image, investigator knowledge of genes and their functions, etc.). Speed then defined "the second simplest microarray data analysis problem: "is it possible to define a differentially expressed gene using multiple [microarray] slides?" These two questions focus on different kinds of variation in microarray data, namely within slide variation, and between slide variation.

Several types of within slide variation are observed for microarray data. Digitized images from a microarray experiment require background correction for optimal data analysis. Several algorithms are available for background correction including SPOT, GenePix, Quantarray and other software. However, not all background correction methods produce the same corrected data values. The Lowess transformation is a procedure that effectively normalizes dye-specific effects from the data, whereas print tip normalization accounts for lack of homogeneity in the array due to print tip effects. Speed developed an extremely useful box plot graphical representation for visualization of inherent array print tip effects.

The amount of variation between microarray slides can differ considerably. It is generally assumed that all slides have the same dynamic range, but this is not observed in practice, so scale normalization is often carried out. Standard deviations are also quite variable for between slide measurements. Several statistical methods are commonly used for evaluating between slide variation, including T-statistic, average and standard deviation. Speed showed that the T-statistic can be driven by sets with a low standard deviation, and averages can be driven by exceptional values (outliers). Thus, if the two methods are applied to the same data, dramatically different outcomes can result. Speed developed a statistic called “B” which is equivalent to an intersection between the T-statistic (low standard deviation) and the average. Speed felt that the “B” statistic may be extremely useful for analyzing between slide variation in microarray datasets.

Panel Discussion*

Cynthia Afshari (Panel Discussion Moderator) opened the afternoon session and commented that the morning session of this workshop focused primarily on analysis of microarray data. She indicated that informatics related to other technologies, such as proteomics, might be discussed during the panel discussion or at a future NCT workshop.

Afshari also pointed out that NCT hopes to develop its CEBS reference database as a useful tool for the entire international research community. The design, establishment and maintenance of this database will present huge technological challenges. The goal of this and the previous two NCT workshops in this series is to solicit input and advice from experts in genomics, proteomics and bioinformatics. Afshari also solicited input from the audience attending the workshop.

The afternoon session began with a short presentation by Paul Spellman (Stanford University) about the Microarray Gene Expression Database Group (http://www.mged.org), which has been working to create a publication standard for microarray data. This group first met in 1999 at the European Bioinformatics Institute, and it was scheduled to have its third meeting in March, 2001 at Stanford University. The goal of MGED is to facilitate adoption of standards for microarray experiment annotation, data representation, experimental controls and data normalization. MGED is doing this so that it will become possible to create useful databases for microarray data.

The MGED Group includes five working groups in the following areas: 1) MIAME – defining the Minimal Information for the Annotation of a Microarray Experiment (MIAME) standard; 2) MAML – developing Microarry Markup Language (MAML), an XML-based mechanism for encoding MIAME compliant data; 3) Data controls and standards; 4) Normalization and quality control; 5) Queries, user groups and data mining.

MIAME specifies data annotation in 6 areas: array, sample, hybridization, controls, authors, and data. Array annotation spells out the construction and characteristics of the microarray chip. Sample annotation defines the biological nature of the sample and the treatment of the sample (*i.e.*, strain, tissue, cell line, dose, time point, etc.). MIAME includes a mechanism for specifying a primary sample and a derived sample. Hybridization annotation specifies the extract protocol and labeling and hybridization procedures. There is currently no agreed upon MIAME standard for data submission, but data format is to be in terms of intensity (and not in terms of a fluorescence ratio).

Panel Discussion Questions

The following 4 questions were given to the panelists and the audience for consideration during the discussion session:

- There are a number of microarray data standardization formats currently under development and concerted efforts to formulate one standard. How best can the development of the NCT database proceed to maintain a specific scope yet remain compatible and flexible with other on-going efforts?

- Microarray and proteomics technologies produce vast amounts of data. The usefulness of a repository of gene expression data is being able to associate it with toxicological and biological information. How best can the NCT initiative integrate biological resources to gain heuristic knowledge and predictive power about chemical effects on biological systems? What considerations should be made for curation?

- What are the highest priorities that the NCT should be considering for research and development in the areas of biostatistics, bioinformatics, and data analysis? How can NIEHS and the NCT best impact and support these areas?

- With the development and implementation of cutting-edge information technology and software tools, where are the fields of data exchange, visualization, and interpretation headed and how best can the NCT leverage such developments to better understand toxicogenomics?

The discussion touched on many of these issues as well as other points of interest. These discussion points are summarized below, grouped according to topic.

Database characteristics

- Afshari asked the panelists if NCT should develop its own database or utilize an existing public database such as GEO. The panelists supported the establishment of an independent database by NCT so that NCT could establish quality assurance filters that might not be used by another database. In addition, NCT could assure that the data deposited in its database would fulfill requirements needed to establish a useful relational and queriable biological database. Quackenbush recommended that the first NCT database structure should be viewed as tentative, and that it should be evaluated and restructured frequently to meet the needs of the NCT and its users.

- Many of the panelists, including Quackenbush, Churchill, Spellman and Wolfinger, stated their opinion that it is essential for the database to include raw data (*i.e.*, intensity values and raw images, when possible). While the image data is invaluable for many reasons, it may be difficult to store large image files in a microarray database. Quackenbush suggested that JPEG files may be useful for storage and reference purposes, although their resolution is insufficient for analytical purposes. There are many image compression methods, and Speed suggested that NCT might consider storing high resolution image files in the database, if they can be compressed sufficiently using one of these methods.

- Panelists repeatedly emphasized that fluorescence intensity values must be stored in the database. They also indicated that the image analysis tools used for data analysis should be recorded in the database.

- In general, panelists agreed that more annotation and information is better than less for microarray data sets. However, the MIAME standards should be adhered to as a minimum. It was also mentioned that the requirements should be sufficiently flexible to account for the fact that different experiments require different approaches and types of information.

- Metrics for data quality were discussed by Yidong Chen (National Human Genome Research Institute). Chen suggested that the fluorescence intensity ratio and the quality of each spot on a microarray should be quantified. He described an approach for measuring spot and array quality using metrics for the size, shape and consistency of a spot and a measurement of local background. Quality values can be assessed for each spot or averaged over a whole array or set of data. Other panelists acknowledged that this is a valuable approach, although it is not yet common practice to analyze both spot intensity and spot quality.

Why contribute to the database?

- Clarice Weinberg (NIEHS) asked the panelists what factors would motivate scientists to contribute their data to a microarray database. The panelists replied to this questions with three points. First, scientists will be motivated to contribute to the database because they are interested in and concerned about scientific progress; scientists will realize that the database will facilitate scientific progress more effectively if more scientists contribute their data to the database. Second, the field is moving towards a consensus concerning standards for microarray data, and scientists will soon be required to meet those standards in order to publish in scientific journals. If the database requirements closely match the publication standard, the process of publishing will require similar annotation and documentation as the process of database submission. Third, the software will be developed to make the processes of data analysis and data submission relatively painless. So the benefit will likely outweigh the cost of contributing to the database.

Standards, Controls and Platforms

- Chris Roberts (Rosetta Inpharmatics, Inc.) discussed use of spike-in samples as controls for data from an inkjet-type oligonucleotide array. He described a protocol that uses a series of spots of one sample at different concentrations. The series is spotted on the array 30 times for a total of 360 standardization spots per array. The same approach can be used for a glass slide array. Roberts and other panelists felt that this approach could be useful for comparing data from different investigators or laboratories, and NCT might consider specifying use of a specific set of standards. Although replicated spiked standards use a significant amount of space on an array, the value of the data set can be increased significantly by including a set of standards.

- Richard Paules (NIEHS) asked if DNA could be used as a standard. Several panelists felt that DNA does not make a useful standard for microarray expression data.

- Several panelists discussed use of an mRNA sample pooled from a group of control animals. For example, an experiment with drug-treated and untreated control animals might use 12 animals per treatment group. It is possible to pool the mRNA from the 12 untreated mice, and assess inter-animal variability by comparing mRNA samples from each treated mouse with the pooled control. This approach was seen as useful; however, it should be kept in mind that the values measured for the pooled sample are effectively an arithmetic average of the values of the individual samples. Churchill advised use of the closest possible control sample whenever possible. An alternate procedure is to use a reference pool and loop design.

- The panelists were asked if the statistical methods presented in the morning session are applicable to data collected on filter arrays and imaged using radiolabeling. All panelists stated that the methods were wholly applicable to this platform. A survey of the audience was taken which revealed that approximately the same number of researchers in the audience use filter arrays or glass slide arrays, and a smaller number use Affymetrix -type arrays.

- The panelists emphasized that experiments should include both experimental replicates and hybridization replicates and that “fluor flip” replicates should be part of the design. They commended the current efforts of the NIEHS Microarray Center, which is already implementing these designs.

NCT priorities

- Afshari asked the panelists what they thought were the highest priority issues that NCT should address. She suggested that there may be concern over the ability of microarray technology to handle low dose or early time point data, because these experiments are among the most difficult to execute. One answer to this question was that it may be important to look at the signal to noise ratio in addition to examining the fluorescence signal per se. Another panelist pointed out that it is important for scientists to use their knowledge of biology to evaluate this kind of data. In some cases, understanding of the biological context can compensate for "noise" in the data. The panelists suggested that in the development of the database there should be integration with other toxicological databases and information.

- Dr. Peter Spencer (Oregon Health Sciences University) emphasized that detailed information about study design and biological outcome should be coupled with microarray data. He emphasized that some effort should be made to identify gene expression fingerprints that are associated with a negative outcome.

Quantification on an absolute scale

- John Frazier (Air Force Research Laboratory) asked the panelists if microarray data can be used to quantify expression on an absolute scale. Frazier argued that absolute quantification is essential in some toxicology studies. The panelists did not agree on the usefulness of an

absolute value measurement of mRNA expression, nor did they agree on whether such a measurement can be produced using microarray data. However, most of the panelists agreed that an absolute value measurement of protein expression is more valuable than an absolute value measurement of mRNA expression.

Quantification, Bioinformatics and Proteomics

- Speed, Churchill, Greller and Chen discussed technical aspects of interpretating microarray data to infer the involvement of specific molecular pathways. The panelists indicated that a variety of statistical and data mining tools are being developed for this purpose. Greller suggested that it is quite feasible to build models based on time variation. Chen emphasized an approach that he has used successfully in which analysis is initiated using a subcluster of genes. Other methods are also applied using an initial unsupervised analysis approach. However, this method of data analysis requires high computational capacity such as a supercomputer and parallel processing.

- Daniel Liebler (University of Arizona) commented on the current state of bioinformatics for proteomics. He noted that it is very difficult to quantify proteomics data and it is also difficult to assess its quality. Proteomics data are very complex, because they convey information on the identity and the modifications of a specific protein species. In addition, repositories for proteomics data are not yet being developed. Liebler encouraged the NCT to promote development of quantitative proteomics, and to try to relate efforts in transcriptome analysis to similar efforts in proteome analysis.

* The panel discussion was moderated by Cynthia Afshari (NIEHS) and the following individuals were panelists:

Pierre Bushel (NIEHS), Yidong Chen, (National Human Genome Research Institute), Gary Churchill (The Jackson Laboratory), Larry Greller (Molecular Mining Corporation), Daniel Liebler, (University of Arizona), Richard Paules (NIEHS), John Quackenbush (The Institute for Genomic Research), Chris Roberts (Rosetta Inpharmatics, Inc.), Terrence Speed (University of California, Berkeley), Paul Spellman (Stanford University), Clarice Weinberg (NIEHS), Russ Wolfinger (SAS Institute, Inc.).